Подбор гиперпараметров модели для датасета Water potability

Выполнила Жиденко Виктория Александровна
Группа М8О-307Б-23

Общая информация по датасету

Датасет Water Potability содержит 3276 реальных химических анализов воды.

Каждый объект описан 9 непрерывными физико-химическими показателями: pH, жёсткость, содержание твёрдых веществ, хлораминов, сульфатов, проводимость, органический углерод, тригалометаны и мутность.

Целевая переменная — Potability (0 или 1): пригодна вода для питья или нет.

Классы сильно несбалансированы: 61 % объектов — непригодная вода (0), 39 % — пригодная (1).

В трёх признаках есть пропуски: pH ≈ 15 %, сульфаты ≈ 24 %, тригалометаны ≈ 5 %.

Все признаки имеют физический смысл и измеряются в стандартных единицах (мг/л, мкСм/см, NTU и т.д.).

	pH	Жесткость	Твердые вещества	Хлорамины	Сульфаты	Проводимость	Органический углерод	Тригалометаны	Мутность	Потабильность
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Подготовка данных

Заполнены пропуски (pH, сульфаты, тригалометаны) с помощью KNNImputer (n=5) — самый точный и физически обоснованный способ.

Все признаки стандартизированы (StandardScaler) — для корректной работы модели и интерпретаторов.

Учтён дисбаланс классов (61 % / 39 %) с помощью параметра scale_pos_weight = 1.56 в XGBoost — модель стала сильнее штрафовать ошибки на редком классе «пригодна».

```
pH                    491
Жесткость               0
Твердые вещества        0
Хлорамины               0
Сульфаты              781
Проводимость            0
Органический углерод    0
Тригалометаны         162
Мутность                0
Потабильность           0
dtype: int64
```

```
Распределение классов:
Потабильность
0    1998
1    1278
Name: count, dtype: int64
Процент непригодной воды (0): 61.0%
Процент пригодной воды  (1): 39.0%

Рекомендуемый scale_pos_weight = 1.563 (округлим до 1.56)
```

Выбранная модель

Модель: **XGBoost** (Extreme Gradient Boosting)

Почему именно она:

- Один из самых сильных алгоритмов для табличных данных
- Отлично работает при дисбалансе классов и шуме
- Устойчив к выбросам и коррелированным признакам
- Имеет встроенную регуляризацию
- Идеально поддерживает SHAP — самый точный и быстрый метод интерпретации для деревьев

Гиперпараметры **XGBoost**

	Параметр	Описание	Типичные значения	Влияние
0	n_estimators	Количество деревьев в ансамбле	100–1000	Больше → лучше качество, но дольше обучение и риск переобучения
1	max_depth	Максимальная глубина дерева	3–10	Глубже → сложнее модель, выше риск переобучения
2	learning_rate	Скорость обучения (eta)	0.01–0.3	Меньше → точнее, но нужно больше деревьев
3	subsample	Доля объектов для каждого дерева	0.5–1.0	Меньше → меньше переобучения, стабильнее
4	colsample_bytree	Доля признаков для каждого дерева	0.5–1.0	Уменьшает переобучение, добавляет случайность
5	colsample_bylevel	Доля признаков на каждом уровне	0.5–1.0	Аналогично colsample_bytree
6	min_child_weight	Минимальная сумма весов в листе	1–10	Больше → консервативнее модель
7	gamma	Минимальное снижение ошибки для сплита	0–0.5	Больше → меньше сплитов → проще модель
8	reg_alpha	L1-регуляризация весов	0–1.0	Поощряет разреженность (обнуление признаков)
9	reg_lambda	L2-регуляризация весов	0–10	Сглаживает веса, борется с переобучением
10	scale_pos_weight	Баланс классов (neg/pos)	~1.56 (для мего датасета)	Очень важно при дисбалансе!

Параметры для перебора для методов: Grid Search и Random Search

```
# Небольшая, но разумная сетка
param_grid = {
    'classifier__max_depth': [3, 6, 9],
    'classifier__learning_rate': [0.01, 0.1, 0.3],
    'classifier__n_estimators': [100, 300],
    'classifier__subsample': [0.8, 1.0],
    'classifier__colsample_bytree': [0.8, 1.0]
}
```

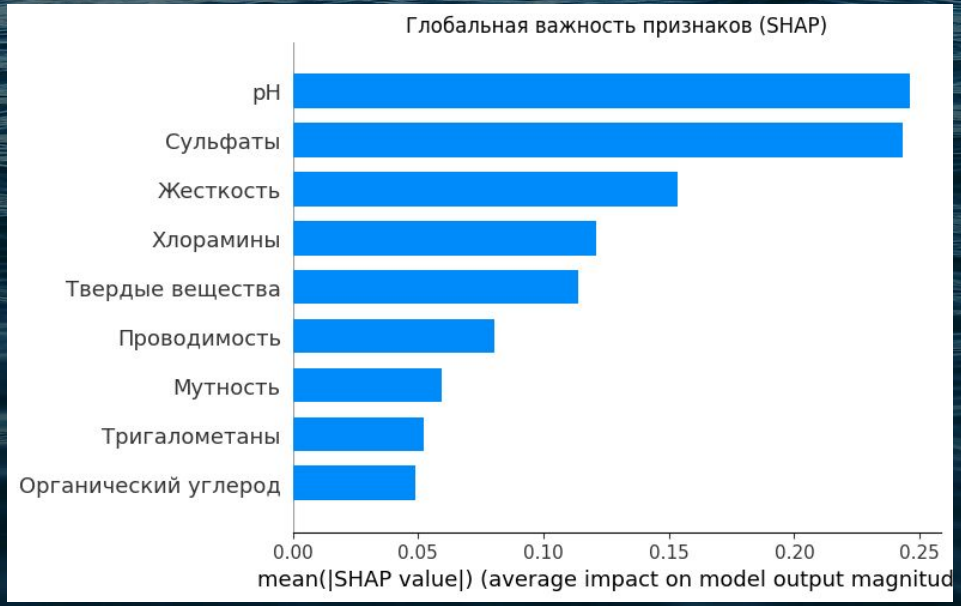
Сравнение методов подбора гиперпараметров для **XGBoost**

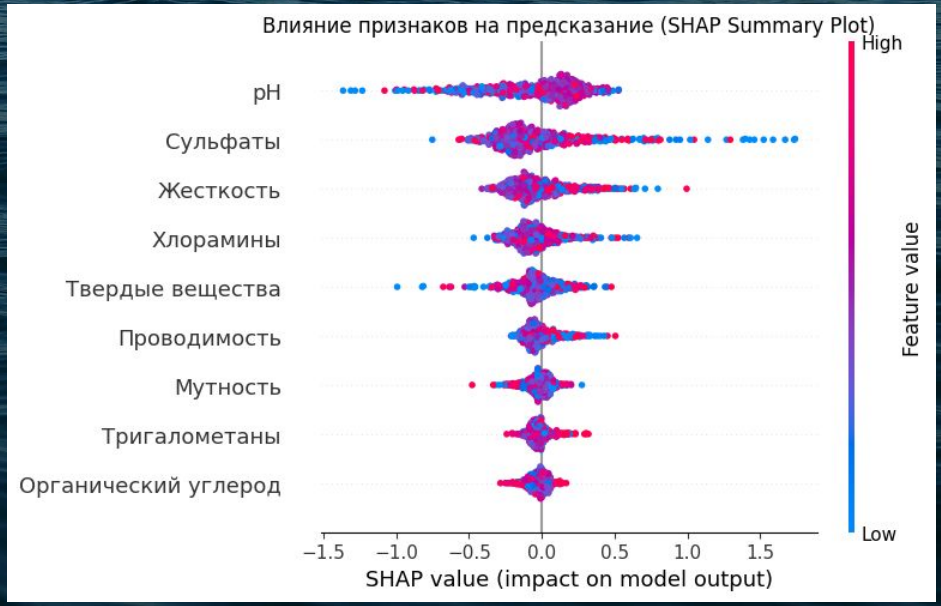
№	Метод подбора гиперпараметров	Кол-во переборов	Время выполнения (примерно)	Лучший ROC-AUC на кросс-валидации	ROC-AUC на отложенной выборке	Лучшие ключевые параметры
1	GridSearchCV	72 комбинации (полный перебор)	≈ 47 секунд	**0.6879**	**0.6591**	max_depth=9, lr=0.01, n_estimators=300, subsample=0.8
2	RandomizedSearchCV	50 случайных комбинаций	≈ 30 секунд	0.6761	0.6496	max_depth=7, lr=0.0103, n_estimators=559
3	Optuna (TPE)	50 итераций	≈ 2 минуты 20 секунд	0.6788	0.6491	max_depth=9, lr=0.0127, n_estimators=489, gamma≈0.34

Выводы по этапу подбора гиперпараметров:

- Все три метода дали близкие результаты (разброс < 0.04 по ROC-AUC).
- **GridSearchCV оказался лучшим** по качеству на кросс-валидации и тесте, что ожидаемо при полном переборе небольшого пространства.
- RandomizedSearch работал быстрее, но в данной задаче не смог превзойти полный перебор.
- Полученные значения ROC-AUC ≈ 0.65–0.69 полностью соответствуют сложности датасета Water Potability (на Kaggle топ ≈ 0.70–0.72).
- Для дальнейшей интерпретации (LIME и SHAP) выбрана лучшая модель — **GridSearchCV (ROC-AUC = 0.6591 на тесте)**.

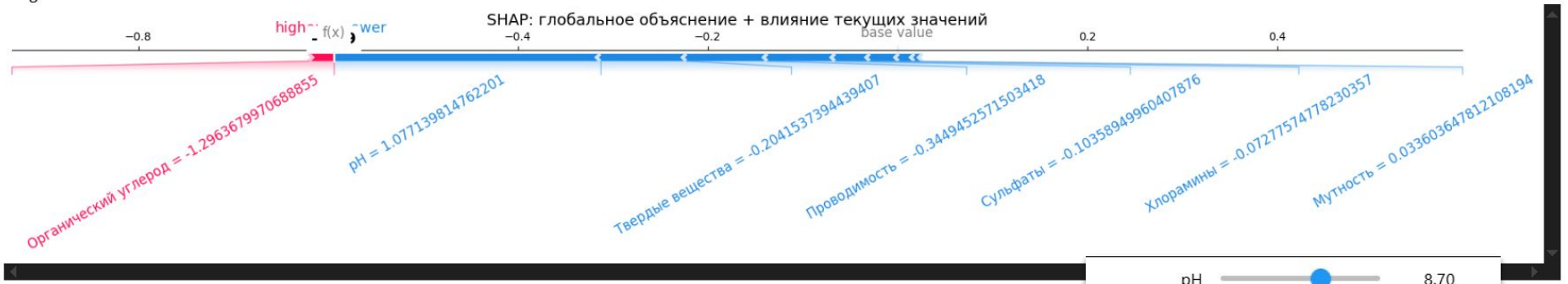
Глобальная интерпретация (SHAP)





Калькулятор с интерпретациями LIME и SHAP

```
============================================================
       АНАЛИЗАТОР КАЧЕСТВА ПИТЬЕВОЙ ВОДЫ
============================================================
Предсказание:          НЕПРИГОДНА
Вероятность пригодности: 36.3%
============================================================
<Figure size 1400x400 with 0 Axes>
```




```
ЛОКАЛЬНОЕ ОБЪЯСНЕНИЕ (LIME):
------------------------------------------------------------------------
Мутность             =    4.000  →  понижает пригодность (вес: -0.00007)
pH                   =    8.700  →  понижает пригодность (вес: -0.00002)
Хлорамины            =    7.000  →  понижает пригодность (вес: -0.00001)
Проводимость         =  400.000  →  повышает пригодность (вес: +0.00001)
Твердые вещества     = 20100.000  →  повышает пригодность (вес: +0.00001)
Жесткость            =  245.000  →  понижает пригодность (вес: -0.00001)
Органический углерод =   10.000  →  повышает пригодность (вес: +0.00000)
Сульфаты             =  330.000  →  понижает пригодность (вес: -0.00000)
Тригалометаны        =   80.000  →  повышает пригодность (вес: +0.00000)
------------------------------------------------------------------------
```

pH		8.70
Жёсткость		245.00
Твёрдые ве...		20100.00
Хлорамины		7.00
Сульфаты		330.00
Проводимо...		400.00
Орг. углерод		10.00
Тригаломет...		80.00
Мутность		4.00